DOCUMENT RESUME

ED 171 756 TM 009 001

AUTHOR Gustafsson, Jan-Eric
TITLE Testing and Obtaining Fit of Data to the Rasch Model.
PUB DATE Apr 79
NOTE 58p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Ability; Complexity Level; *Goodness of Fit; *Item Analysis; *Mathematical Models; Predictive Measurement; *Research Problems; Scores; Statistical Bias; Statistical Studies; Test Items
IDENTIFIERS Maximum Likelihood Estimation; *Rasch Model

ABSTRACT

Problems and procedures in assessing and obtaining fit of data to the Rasch model are treated and assumptions embodied in the Rasch model are made explicit. It is concluded that statistical tests are needed which are sensitive to deviations so that more than one item parameter would be needed for each item, and more than one person parameter would be needed for each person. Statistical goodness-of-fit tests--based on the conditional maximum likelihood estimates of the item parameters--which can detect these two kinds of deviation are presented. Common sources of deviation are also identified, as are the tests needed to detect them. Problems in the use of statistical tests to assess fit are discussed and some investigations of power are presented. In relation to a distinction between use of the Rasch model as a criterion and as an instrument, the treatment of the goodness-of-fit problem in different measurement contexts is discussed. Finally, it is concluded that items which can be identified as misfitting should not be routinely excluded to obtain fit to the model; instead other actions should often be taken--such as grouping of the items into homogeneous subsets. (Author/CP)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED171756

# TESTING AND OBTAINING FIT OF DATA TO THE RASCH MODEL 1)

Jan-Eric Gustafsson
University of Göteborg

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jan-Eric Gustafsson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 8-12, 1979. Requests for copies should be sent to Jan-Eric Gustafsson, Institute of Education, University of Göteborg, Fack, S-431 20 Mölndal, Sweden.

TM009 001

1) The research presented in this paper has been supported financially by the Swedish Council for Research in the Humanities and the Social Sciences and by the National Board of Education.

ABSTRACT

Problems and procedures in assessing and obtaining fit of data to the Rasch model are treated in the paper. The assumptions embodied in the model are made explicit and it is concluded that statistical tests are needed which are sensitive to deviations such that more than one item parameter would be needed for each item, and such that more than one person parameter would be needed for each person. Statistical goodness-of-fit tests, based on the conditional maximum likelihood estimates of the item parameters, which can detect these two kinds of deviation are presented. Common sources of deviation are also identified, as are the tests needed to detect them. Problems in the use of statistical tests to assess fit are discussed and some investigations of power are presented. In relation to a distinction between use of the Rasch model as a criterion and as an instrument the treatment of the goodness-of-fit problem in different measurement contexts is discussed. Finally it is concluded that items which can be identified as misfitting should not be routinely excluded to obtain fit to the model; instead other actions should often be taken such as grouping of the items into homogeneous subsets.

## Introduction

Thecrists and practicioners are to an increasing extent focussing attention on what is called the latent trait (LT) models within test theory (Baker, 1977; Hambleton & Cook, 1977). The LT models specify a relationship between observable examinee performance and an unobservable trait assumed to underlie performance. Their great power stems from the fact that parameters describing characteristics of the test items can be estimated in such a way that they are invariant from one group of persons to another, and estimates of the ability of persons can be made in such a way that they are invariant from one sample of items to another.

The family of LT models has many members (Lord & Novick, 1968; Hambleton, Swaminathan, Cook, Eignor & Gifford, 1977) but the most important ones seem to be certain models for dichotomous items, based on logistic functions. The simplest of these is the Rasch model (Rasch, 1960, 1966), or the 1-parameter logistic model. In the Rasch model one parameter only is used to describe each item, but there are also other models such as the 2- and 3- parameter models (Birnbaum, 1968), in which additional parameters are used to describe characteristics of the items.

The Rasch model has important theoretical and practical advantages when it comes to the estimation of parameters (Andersen, 1973a; Fischer, 1974; Gustafsson, 1977). The relative simplicity of the Rasch model also makes it easy to apply the model in solving practical measurement problems, such as linking and equating tests, optimizing tests, carrying out tailored testing, constructing item banks and so on (c.f. Wright, 1977a). These reasons are sufficient to explain why the Rasch model is the LT model most frequently applied.

The LT models have very desirable characteristics which make possible the solution of measurement problems which are difficult or impossible to solve within the framework of classical test theory. But the models entail strong assumptions about the nature of the data, and unless these assumptions are fulfilled, the validity of the results of applications is endangered. The Rasch model is the most constrained one, and

it is also the model which entails the strongest assumptions. The question of whether the data fit the models or not is therefore of great importance.

More specifically, there are three reasons why the question of fit is an important one. In the first place, it is important to realize that if the assumptions are fulfilled for a set of data, then all the desirable characteristics of the LT models are logical implications of the mathematical structure of the models themselves; the validity of applications need therefore not be empirically proven if the data fit a model. Secondly, in some cases fit to a model is an important end in itself, because the models, and above all the Rasch model, formalize desirable characteristics of measurements (cf. Gustafsson, 1977; Wright, 1977b). Thirdly, in those cases where, for some reason, it is necessary to use an LT model without the data fitting it, it is essential that the deviations from the model are reasonably well-known, since different applications are endangered to different degrees depending on the type of deviation.

This paper deals with the problems of assessing and obtaining fit of data to the Rasch model. This model is concentrated upon because of the advantages it has over other models, and because it entails the strongest assumptions.

Ever since the model was first formulated by Rasch (1960) the problem of fit has been studied, and statistical tests of goodness-of-fit have been developed (Wright & Panchapakesan, 1969; Andersen, 1973b; Martin-Löf, 1973; Fischer, 1974; Mead, 1976a, 1976b). But there are factors which motivate another treatise on the subject.

The development of computational algorithms (Gustafsson, 1977, 1979) has made another class of statistical tests of goodness-of-fit available for general use. These are based on the conditional maximum likelihood approach to estimation of item parameters in the Rasch model (Andersen, 1973b; Martin-Löf, 1973), and they have better statistical properties than most other goodness-of-fit tests. Even more important, however, is

the fact that there are such tests, only described in an unpublished report by Martin-Löf (1973), which are sensitive to deviations from the model that are difficult to detect with other methods. These statistical tests are presented in the paper, along with a presentation of the conditional approach to estimating the item parameters.

The sensitivity of different statistical tests to different sources of deviation from the Rasch model has not been much studied, and an attempt is made to shed light upon this problem. A few studies of the power of the goodness-of-fit tests as a function of factors such as sample size and number of items are also presented.

Closely associated with the problem of fit is the question of the robustness of the model. Analyses of that problem are presented in relation to one particular kind of application, the equating of tests.

Strategies used to obtain fit of data to the model are also discussed and problems inherent in the most commonly used strategy are identified. On the basis of that criticism an alternative strategy is outlined.

## 1. The Rasch model and its assumptions

According to the Rasch model, the probability of a correct answer to an item is a function of two parameters only, one describing the difficulty of the item ($\sigma_i$, i=1,...,k) and one describing the ability of the person taking the test ($\xi_v$, v=1,...,n). If we denote a correct answer to item i by person v as $A_{vi}=1$ the probability of this outcome is:

$$(1.1) \qquad P(A_{vi}=1 \mid \xi_v, \sigma_i) = \frac{\exp(\xi_v - \sigma_i)}{1 + \exp(\xi_v - \sigma_i)}$$

The item characteristic curve (ICC) is a central concept in LT theory. The ICC is the function relating the probability of a correct answer to an item (i) to the ability variable ($\xi$). From (1.1) follows that the ICC for an item in

the Rasch model is:

$$(1.2) \qquad f_i(\xi) = \frac{\exp(\xi-\sigma_i)}{1 + \exp(\xi-\sigma_i)}$$

The ICC is a function of one parameter only, the item parameter, or a term which will be used interchangeably, the difficulty.

In the Rasch model unidimensionality is assumed since there is only one parameter of ability. However, we will need a more exact definition of unidimensionality. Lord and Novick (1968) presented two definitions of unidimensionality in LT models.

One of their definitions (p. 359) is actually a definition of dimensionality of any order but here it has, along with some changes in notation, been rewritten as a definition of unidimensionality:

> Consider a set of k items and one latent trait $\xi$ which affects examinee performance on all items in the set. We can now represent each examinee as a point on the trait. Next, consider all the examinee populations that may be of interest for this set of k items. Assume that each item is administered just once to each examinee, and consider the conditional frequency distribution (over people) of item score for any fixed value of $\xi$. If this (unobservable) distribution is not the same for all the populations of examinees, then there must be one or more psychological dimension in addition to $\xi$, that discriminate among the populations of interest. In defining the complete latent space, therefore, we must include these additional dimensions. Thus, by definition, in the complete latent space the conditional distribution of item score for fixed $\xi$ is the same for all populations of interest.

From this definition of unidimensionality follows that the ICC for an item is invariant for those populations used to define the complete latent space.

The other definition of unidimensionality given by Lord & Novick (1968) is founded on the concept of local (or conditional) statistical independence. If we use the algebraic notation $A_{vi}=a_{vi}$ to represent the score, 0 or 1, of person v to item i, the assumption of local statistical independence

can be written:

$$(1.3) \qquad P(A_{v1}=a_{v1}, A_{v2}=a_{v2}, \ldots, A_{vk}=a_{vk} | \xi_v) = \prod_{i=1}^{k} P(A_{vi}=a_{vi} | \xi)$$

Thus, if local independence holds the probability of an examinee response pattern is given by the product of the probabilities of the item responses, and the Lord and Novick (1968, p.540) definition says that if (1.3) holds for some real-valued trait $\xi$, the measurements satisfy a unidimensional latent-trait model.
This abstract definition is given a more concrete meaning when formulated as follows:

> "...an individual's performance depends on a single underlying trait if, given his value on that trait, nothing further can be learned from him that can contribute to the explanation of his performance. The proposition is that the latent trait is the only important factor and, once a person's value on the trait is determined, the behavior is random, in the sense of statistical independence" (Lord & Novick, 1968, p. 538).

The difference between the two definitions of unidimensionality is that the latter one explicitly introduces the assumption of local statistical independence. However, there is no conflict between the two definitions since the assumption of local statistical independence is equivalent to the assumption that the latent variable under consideration spans the complete latent space (Lord & Novick, 1968, p. 361).

It is also necessary to consider another attempt to define unidimensionality. Lumsden (1978) formulated a statistical model in which both items and persons are located on an attribute continuum (latent variable). In contrast to the Lord & Novick approach the items, and not the persons, have a point location on the continuum, while the persons are assumed to have a distribution of attribute locations, resulting from moment to moment fluctuations. The distribution of attribute locations is not assumed to be the same for all persons, taking into account the possibility that persons may differ in reliability (Lumsden, 1977).

In the Lumsden formulation, unidimensionality of the items is assured by the fact that they are located on the same attribute

continuum. However, the Lumsden model is not unidimensional in the sense of Lord and Novick's definition, which is best seen if the Lumsden formulation is formalized. One way to do this is to generalize the Rasch model, taking into account the possibility of varying person reliabilities, by adding another parameter for each person, ($\psi_v$, v=1,...,n), which, in accordance with Mead (1976b), will be referred to as the sensitivity parameter:

$$(1.4) \qquad P(A_{vi}= 1|\xi_v,\psi_v,\sigma_i) = \frac{\exp\psi_v(\xi_v-\sigma_i)}{1+ \exp\psi_v(\xi_v-\sigma_i)}$$

This model will be referred to as the Lumsden model. In this model the PCC's (person characteristic curves) in which for each person the probability of a correct answer is shown as a function of item difficulty, are not parallel , with the sensitivity parameter reflecting the slope of the PCC.

In the Lumsden model knowledge of a person's ability parameter $\xi_v$, which can be interpreted as the mean of his distribution of attribute locations, could not alone explain his performance, since the sensitivity parameter would also be needed. Therefore, the Lumsden model is not unidimensional in the sense of Lord and Novick's definition of the term.

It is in all likelihood impossible to obtain separate estimates of the $\xi_v$ and the $\psi_v$ parameters, so the Lumsden model is not useful as an LT-model. It is useful, however, as an alternative model to the Rasch model in investigations of fit since it does specify a certain kind of multidimensionality.

Another assumption in the Rasch model is that all the items have the same discriminative power, i.e. that all the ICC's are parallel. The meaning of this assumption is most clearly seen if the Rasch model is contrasted with the Birnbaum (1968) model, or the 2- parameter model which introduces another parameter for each item, the discrimination parameter ($\alpha_i$,i=1,...,k). According to the Birnbaum model, the ICC for an item is:

$$(1.5) \qquad f_i(\xi) = \frac{\exp\alpha_i(\xi-\sigma_i)}{1+\exp\alpha_i(\xi-\sigma_i)}$$

The discrimination parameter reflects the relation between performance on an item and the latent variable, and if the discrimination parameter is different among the items in a set, the ICC's will be non-parallel.

The discrimination parameters are different if the items in addition to the common latent trait reflect different "specific" factors and/or if they are differently affected by random errors. However, it is difficult to define a specific factor as opposed to another latent variable, of which Lord and Novick (1968) were aware:

> The psychometrician is likely to wish to define his complete latent space to include all "important" psychological dimensions that affect performance on a given set of items and to exclude those variables that comprise "errors of measurement". Unfortunately, it seems logically impossible to distinguish objectively those variables that are simply "errors of measurement" from those that are not (p. 340).

Thus, even at the level of formal definition it is difficult to make a distinction between the assumption of unidimensionality and the assumption of homogeneous item discrimination.

The specific factors are assumed to be uncorrelated with the ability measured by all the items and with the specific components of all the other items. However, even though the specific factors can fulfill the assumptions of orthogonality when we confine our attention to a specific sample of items, they are not likely to do so in the "population of items" (cf. Lumsden, 1978). Since generalization is practically always intended beyond a certain set of items, the distinction between the assumption of unidimensionality and the assumption of homogeneous item discrimination becomes even more difficult to uphold.

Three assumptions in the Rasch model have been discussed: the assumption of unidimensionality, the assumption of local statistical independence, and the assumption of homogeneous item descrimination; these are also the assumptions commonly associated with the Rasch model (cf. Gustafsson, 1977; Hambleton et al.,1977).

It has been concluded, however, that the assumption of unidimensionality and the assumption of local statistical independence are either identical, or inseparable, and also that it is difficult to uphold any clear distinction between the assumption of homogeneous item descrimination and the assumption of unidimensionality.

It does seem that the Rasch model assumptions can be violated in basically two ways: either a model is needed to describe the data which contains two or more parameters for each person, which would be a violation of the assumption of unidimensionality; or a model is needed which contains two or more parameters for each item, which would be a violation of the assumption of the form of the ICC's; or, of course, a combination of these.

If the Rasch model holds true for a set of data the item parameters are invariant from one group of persons to another and the person parameters are invariant from one group of item to another. But if more than one parameter is needed for each person, such as is the case in the Lumsden model, for example, the person parameters will not be invariant for groups of items. If more than one parameter is needed for each item, such as is the case in the Birnbaum model, for example, the item parameters will not be invariant for groups of persons. This forms the basic rationale for the statistical methods of investigating fit to the Rasch model. The statistical tests will be taken up later on, after the basics of the conditional maximum likelihood approach to estimating the item parameters in the Rasch model have been presented.

## 2. The conditional approach to the Rasch model

There are several different approaches, ranging in mathematical and statistical sophistication, to the problem of estimating the parameters in the Rasch model from a set of observational data. There are, for example, simple methods suited for hand calculations (e.g. Wright & Douglas, 1975). But these methods introduce further assumptions, such as an assumption of normality of the distribution of person parameters, and these methods are only approximate. When the user of the model has access to a computer, better methods of estimation become available.

The most commonly used methods of estimation are based on the maximum likelihood approach. However, two entirely different maximum likelihood estimators have been defined for the item parameters in the Rasch model. One is what is called the unconditional maximum likelihood (UML) approach in which the item parameters and the person parameters are estimated simultaneously. (Wright & Pachapakesan, 1969; Wright & Douglas, 1977). The other is what is called the conditional maximum likelihood (CML) approach, in which the likelihood function for estimating the item parameters is expressed in the item parameters only, through conditioning on raw score (Andersen, 1973a; Fischer, 1974; Gustafsson, 1977).Only in the Rasch model is this possible, because raw score is a sufficient estimator of the person parameter.

Only the CML estimator yields consistent estimates of the item parameters (Andersen, 1973a; Fischer, 1974), but the UML estimator is the one most commonly used (Wright & Douglas, 1977). There are two reasons why the theoretically inferior UML estimator has been used instead of the CML estimator. In the first place, the CML estimates are computationally more cumbersome than the UML estimates and they have even been impossible to compute for anything but short tests. Secondly, it has been shown (Wright & Douglas, 1977) that if a correction is made of the UML estimates, they come close to the CML estimates.

If the similarity between the estimates obtained with the UML and CML approaches was the only issue in the choice between the two approaches, there would be little reason to use the CML approach. There is, however, another, more important difference. On the basis of the CML approach it is possible to devise efficient statistical tests of fit with known statistical properties, while under the UML approach only approximate statistical tests have been formulated.

The computational problems in relation to the CML algorithm have recently been solved (Gustafsson, 1977, 1979) so that now the CML estimates can be obtained for long tests (80-100 items, say) as well, and most often with a relatively limited amount of computational work. Therefore only the CML approach will be

considered in detail in the sequel of this paper.

The mathematical notation becomes greatly simplified if an antilogarithmic transformation is made of the parameters in the Rasch model such that $\theta_v=\exp(\xi_v)$ and $\varepsilon_i=\exp(-\sigma_i)$. The probability of observing the outcome $A_{vi}=a_{vi}$ can then be written:

$$(2.1) \qquad P(A_{vi}=a_{vi}|\theta_v,\varepsilon_i)=\frac{(\theta_v\varepsilon_i)^{a_{vi}}}{1+\theta_v\varepsilon_i}$$

We want to estimate the parameters from the answers of n persons to k items, and assemble the scores into the matrix $((a_{vi}))$. The raw score for person v is:

$$(2.2) \qquad r_v=\sum_{i=1}^{k} a_{vi}$$

and the total number of correct responses to item i (the item score) is:

$$(2.3) \qquad s_i=\sum_{v=1}^{n} a_{vi}$$

Those persons who have 0 or k correct answers must be excluded from the $((a_{vi}))$ matrix since no estimates of their parameters can be obtained and items with 0 or n correct answers must be excluded for the same reason.

Consider first a given examinee with raw score $r_v$ and person parameter $\theta_v$. Given a set of items with parameters $(\varepsilon_i)$, the probability that this examinee obtains any score vector $(a_{vi})$, assuming independence of the responses, is:

$$(2.4) \qquad P\{(a_{vi})|\theta_v,(\varepsilon_i)\}=\prod_{i=1}^{k}\frac{(\theta_v\varepsilon_i)^{a_{vi}}}{1+\theta_v\varepsilon_i}=\frac{\theta_v^{r_v}\prod_i\varepsilon_i^{a_{vi}}}{\prod_i(1+\theta_v\varepsilon_i)}$$

To be able to express this probability as a conditional probability, given score $r_v$, we must know the probability of obtaining score $r_v$ given $\theta_v$. This latter probability is given by the sum of the probabilities of all possible ways of obtaining the

score $r_v$, that is the sum of all the expressions such as (2.4) in which the vector $(a_{vi})$ sums up to r.

A given score r obtained on k items can of course be obtained in $\binom{k}{r}$ different ways. We will need a special notation to be able to express this in a simple way. Define:

$$\gamma_r\{(\varepsilon_i)\} = \sum_{\Sigma_i a_{vi}=r} \prod_{i=1}^{k} \varepsilon_i^{a_{vi}} \tag{2.5}$$

The $\gamma_r\{(\varepsilon)\}$ (or, for short, $\gamma_r$) is called the <u>elementary symmetric function of order r in the parameters $(\varepsilon_i)$</u>.

We can now write the probability of obtaining the score r, given $\theta_v$ and $(\varepsilon_i)$:

$$P\{r|\theta_v,(\varepsilon_i)\} = \sum_{\Sigma_i a_{vi}=r} \prod_{i=1}^{k} \frac{(\theta_v\varepsilon_i)^{a_{vi}}}{1+\theta_v\varepsilon_i} = \frac{\theta_v^r \gamma_r}{\prod_i (1+\theta_v\varepsilon_i)} \tag{2.6}$$

Thus the conditional probability of obtaining any vector $(a_{vi})$ with the total score $r_v$, given the score $r_v$, is:

$$P\{(a_{vi})|r,(\varepsilon_i)\} = \frac{P\{(a_{vi})|\theta_v,(\varepsilon_i)\}}{P\{r|\theta_v,(\varepsilon_i)\}} = \frac{\prod_{i=1}^{k} \varepsilon_i^{a_{vi}}}{\gamma_r} \tag{2.7}$$

If independence is assumed between examinees, the conditional likelihood of the data matrix $((a_{vi}))$ is easily obtained. The logarithm of the likelihood function can be shown (Fischer, 1974; Gustafsson, 1977; Wright & Douglas, 1977) to be:

$$\log \Lambda = \sum_{i=1}^{k} s_i \log\varepsilon_i - \sum_{r=1}^{k-1} n_r \log\gamma_r \tag{2.8}$$

where $n_r$ is the number of persons with raw score r.

Estimation equations for the item parameters can be derived from (2.8) (Fischer, 1974; Gustafsson, 1977; Wright & Douglas, 1977). The greatest problem in solving the equations, which must be done iteratively, lies in efficiently and accurately

computing the $\gamma_r$ and their first derivatives with respect to each of the items ($\gamma_{r-1}^{(i)}$) and sometimes also their second derivatives with respect to the items two at a time ($\gamma_{r-2}^{(i,j)}$). However, as was shown by Gustafsson (1977, 1979) it is possible to devise recursive formulas that can manage these tasks.

It is not necessary to treat methods for estimating the person parameters since it is possible to avoid estimation of these in evaluations of fit. This is quite fortunate since the statistically correct method of estimating the person parameters is impossible to apply in practical work (Fischer, 1974, pp. 239-240).

## 3. Goodness-of-fit tests for the Rasch model

All the goodness-of-fit tests are based on the principle that implications of the model assumptions are tested against observable results. But there are several implications of the model assumptions and the tests can technically and statistically be constructed in many different ways, so there are several goodness-of-fit tests for the Rasch model.

Rasch (1960, 1966) showed that it is possible to devise a test of the model in which no use is made of estimated parameters. This test, which is a generalization of Fisher's exact test for a 2x2 matrix is, however, computationally so cumbersome that it has as yet proven impossible to put it into practical use. Therefore, all the goodness-of-fit tests in practical use, employ estimates of parameters in the model, and tests based on the UML- and CML-approaches differ greatly.

Wright and Panchapakesan (1969) developed within the framework of the UML-approach, a test of overall fit and a test of item fit based on comparisons between observed and theoretically expected frequencies of correct answers to each item at different levels of ability. Mead (1976a, 1976b) extended this approach into a method based on analysis of residuals in the fitted model, using analysis of variance procedures and plots of the residuals. This procedure allows detection of different types of deviation from the model, such as guessing,

speededness and learning effects.

The distributions of the test-statistics formulated within the framework of the UML-approach are unknown, however. The chi-square and z-distributions have been relied upon, but simulation studies indicate that even though the means of the distributions conform to the expected ones, the variances may depart substantially (Mead, 1976b).

Within the framework of the CML-approach goodness-of-fit tests have been formulated (Andersen, 1973b; Martin-Löf, 1973) which have at least asymptotically known distributions, and which have been shown to be parametric counterparts to Fisher's exact test (Martin-Löf, 1974b). These tests are presented below.

Tests sensitive to variations in the ICC's

It has already been concluded that if a set of data fit the Rasch model, the item-parameters (or the ICC's) will be invariant for groups of persons. Andersen (1973b ;cf Martin-Löf, 1973) has presented a conditional likelihood ratio test of model fit from this starting point.

To compute this test the item parameters are estimated using the total sample of persons, and also within g disjoint subgroups of persons with $n_j$ (j=1,...,g) persons in each. In each estimation of the item parameters a maximum of the logarithm of the likelihood (2.8) is obtained. We can call the maximum obtained for the total group of persons $H_t$ and the maxima obtained for the subgroups $H_j$ (j=1,...,g). The following test statistic can then be written:

$$\log\lambda = H_t - \sum_{j=1}^{g} H_j \qquad (3.1)$$

Andersen (1973b) has shown that $-2\log\lambda$ is asymptotically chi-square distributed with (g-1)(k-1) degrees of freedom when each $n_j \to \infty$.

This test is sensitive to differences in the ICC's for different groups of persons and will therefore be referred to as the A-ICC test. However, the persons can be grouped according to different criteria, and depending upon how the grouping is done the test is sensitive to different violations of the model assumptions. One possibility is to group persons according to level of performance on the test, i.e. according to raw score. When used in this way the test is sensitive to variations in the slopes of the ICC's, i.e. it guards against the alternative hypothesis that the Birnbaum model, or a model with even more parameters for each item, would be needed to describe the data (cf. Andersen, 1973b). We will use a special name of the test for this important kind of application: the A-ICCSL test, with the postfix SL chosen to indicate that the test investigates the homogeneity of the slopes of the ICC's.

But the persons can also be grouped according to other criteria such as sex, social background, or school, just to mention a few. When used in this way the A-ICC test is a test of unidimensionality since it follows directly from the definition of unidimensionality that the ICC for an item must be invariant for groups of persons. This holds true in particular when the grouping is not confounded with level of performance since then the test would also be sensitive to variations in the slopes of the ICC's.

Martin-Löf (1973, pp. 128-129) has suggested another test which is sensitive to variations in the slopes of the ICC's and it will be referred to as the ML-ICCSL test. This test is asymptotically equivalent with the A-ICCSL test but it is of quite a different construction. In the ML-ICCSL test the item parameters are only estimated for the total group, and the test is computed from the differences between observed and predicted frequencies of correct answers for persons with different raw scores (score groups).

Let $n_{ir}$ denote the observed frequency of correct answers to item i for those persons who have r correct answers. A corresponding predicted frequency can also be determined: The conditional probability that a person with raw score r answers item i correctly

can easily be shown to be $\frac{\varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r}$. Therefore, if the model is true for the data, the following relationship should hold approximately true:

$$n_{ir} \simeq \frac{n_r \varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r} \qquad (3.2)$$

The ML-ICCSL test takes as its starting point this relationship and from the deviations between observed and predicted frequencies a chi-square sum is built up.

If we label the vector $\begin{pmatrix} n_{1r} \\ \vdots \\ n_{kr} \end{pmatrix} = (q_r)$ and call the corresponding vector of predicted frequencies $\begin{pmatrix} \frac{n_r \varepsilon_1 \gamma_{r-1}^{(1)}}{\gamma_r} \\ \cdot \\ \cdot \\ \cdot \\ \frac{n_r \varepsilon_k \gamma_{r-1}^{(k)}}{\gamma_r} \end{pmatrix} = (t_r)$ the test statistics is:

$$T = \sum_{r=1}^{k-1} \{(q_r)-(t_r)\}' \{((V_r))\}^{-1} \{(q_r)-(t_r)\} \qquad (3.3)$$

in which quadratic form $((V_r))$ is a variance- covariance matrix of order kxk with elements defined as follows:

$$\begin{cases} \frac{n_r \varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r} & \text{in the diagonal} \\ \\ \frac{n_r \varepsilon_i \varepsilon_j \gamma_{r-2}^{(i,j)}}{\gamma_r} & \text{for } i \neq j \end{cases} \qquad (3.4)$$

Martin-Löf (1973) has shown that the test statistic is asymptotically chi-square distributed with (k-1)(k-2) degrees of freedom when each $n_r \to \infty$.

In (3.3) the summation is made over all score groups. If, however, some $n_r=0$ we have to restrict the summation to those R groups in which $n_r>0$. The degrees of freedom then are (k-1)(R-1).

When k is large, the test is quite tedious to compute since it requires computation of k-1 matrix inversions as well as the second derivatives of the symmetric functions. It can be noted, however, that the actual inversion of the matrices can be avoided: Scheffé (1959) has shown that the quadratic form can be computed by evaluating two determinants instead, which requires less computational work.

The ICCSL tests give information about the homogeneity of the slopes of the ICC's, but they do not give any information of value concerning the reasons for poor fit. Due to the lack of a statistical test of item fit with a known distribution under the CML-approach, graphical methods have been resorted to. This is no great sacrifice, however, since the logic of testing the fit of single items can be questioned (see section 7 below), and since descriptive information is needed more than anything else.

The relationship (3.2) can be rewritten so that it expresses a relationship between proportions of correct answers, instead of frequencies. If, for a fixed item, the observed proportion is plotted against the predicted proportion, the points for the score groups should fall along a straight line with a slope of unity, even though the points as a function of stochastic variation will be spread around the line of unit slope. This graphical test will be referred to as the GR-ICCSL test, since it is sensitive to variations in the slope of the ICC's.

The plots that are observed in applications of the model tend to have many different appearances. However, 3 different types account for the absolute majority of the patterns observed. The first is where the points actually fall close to the line of unit slope, and this indicates fit to the model. The second

type of pattern appears when the observed proportion of correct answers for the lower score groups is higher than the predicted proportion, while at the same time, the observed proportion is lower than the predicted proportion for the higher score groups. A low discrimination parameter would be found for such an item if the Birnbaum model was applied. The third pattern, finally, appears when the observed proportion for the lower score groups is lower than the predicted one and when the observed proportion for the higher score groups is higher than the predicted one, and it reflects the case when the item has too high a discrimination parameter.

Test sensitive to variations in the PCC's

The tests presented above all investigate the invariance of the item parameters for groups of persons and they will therefore be referred to as ICC-tests. Practically all other tests of fit which have been used also belong to the group of ICC-tests and in particular to the sub-group of ICCSL-tests. It is easily shown, however, that there may be violations of the assumptions of the Rasch model which cannot be detected with these tests. Lumsden (1978), for example, showed that the PCC's may be non-parallel, while the ICC's are parallel.

To investigate the hypothesis that the Lumsden model, or another model with more than one person parameter is in fact needed to represent the observations, one could study the invariance of the person parameters for groups of items. However, a test constructed straightforwardly from this point of departure would have less than optimal characteristics, since a very large number of parameters would have to be estimated, and since it is practically impossible to estimate the abilities conditionally on item score.

It is, however, not necessary to estimate the abilities to perform the test. A conditional likelihood ratio test , founded on the CML estimates of the item parameters, which tests the hypothesis that two groups of items measure the same ability has been presented by Martin-Löf (1973, pp. 135-136; cf. Leunbach, 1976).

To compute the test it is necessary that the items be grouped into two disjoint sets. Let us say that there are $k_1$ and $k_2$ items in the two sets, respectively, and that $k_1+k_2=k$. Furthermore, let $n_{r_1r_2}$ be the number of persons with raw score $r_1$ on the first set and raw score $r_2$ on the second set. When the item parameters for the total set of k items are estimated, a maximum of the logarithm of the likelihood function (2.8) is obtained ($H_t$) and when the item parameters are estimated for each set separately, the corresponding maxima $H_1$ and $H_2$ are obtained. The following test statistic can then be formed:

$$(3.5) \qquad \log\lambda = -\sum_{r_1=0}^{k_1} \sum_{r_2=0}^{k_2} n_{r_1r_2} \log \frac{n_{r_1} n_{r_2}}{n} + \sum_{r=0}^{k} n_r \log \frac{n_r}{n} + H_t - H_1 - H_2$$

Martin-Löf (1973) has shown that $-2\log\lambda$ is approximately chi-square distributed with $k_1k_2-1$ degrees of freedom when $n \to \infty$.

The test can be applied with the items grouped according to different principles, and depending upon how the items are grouped the test will be sensitive to different violations of the assumptions. One possibility is to group the items according to item score, i.e. difficulty. Then the test investigates the hypothesis that a model of the Lumsden type would be needed to account for the observations, i.e. that the person sensitivity parameters differ. In this special kind of application the test will be referred to as the ML-PCCSL test, since it tests the homogeneity of the slopes of the PCC's.

But the test can also be applied with the items grouped according to different hypothesized dimensions. In this kind of application the test is, of course, a direct test of unidimensionality, and when used in this way it will be referred to as the ML-PCC test.

It should also be pointed out that the test will also be sensitive to a difference in the mean value of the discrimination parameter for the two sets of items. Within the sets of items the discriminations can vary, however, without this being detected by the test as long as the mean discrimination is the same.

With the possibility of varying person sensitivity parameters in mind, the question of person fit to the model is actualized. Under the CML approach it is at least theoretically simple to construct a test of person fit.

An expression has already been derived for the probability of obtaining any given score vector, given a certain raw score (2.7). A p-value is obtained if the probability of all more extreme score vectors, i.e. those with a lower or the same conditional probability of being observed, is summed up. Unfortunately this test is computationally cumbersome since even with few items the total number of possible score vectors is very large.

A computationally more feasible test can be constructed if the items are grouped into sets. Consider the case when only two sets of items are used, with $k_1$ and $k_2$ items. Let us, for any given person, denote the raw score on the first set $r_1$ and the raw score on the second set $r_2$, with $r_1+r_2=r$. Denote further the symmetric functions of the corresponding orders in the item parameters, estimated with both sets pooled, as $\gamma_{r_1:1}$ and $\gamma_{r_2:2}$, respectively. It is then easily shown (cf. Leunbach, 1976) that the conditional probability of obtaining the raw scores $r_1$ and $r_2$ is:

$$(3.6)\quad P(r_1|r) = \frac{\gamma_{r_1:1}\ \gamma_{r_2:2}}{\gamma_r}$$

A p-value for the fit of the person is obtained if the probabilities of all equally or more extreme combinations of raw scores on the groups of items are summed up.

A test of this kind is easy to compute. It can be suspected to have a low power, however, and the power would also be very different for different raw scores if the same grouping of items is used. Power can be increased however, if the test is generalized to more than two groups of items and if a different grouping is used for each raw score.

It must be pointed out that a test like this cannot be applied to all the persons in a sample, since the significance level would then be seriously disturbed. Only when a single randomly chosen person is observed does a statistical test of person fit have any meaning.

## 4. Sources of deviation from the Rasch model -- and how they are detected

In the previous section we have seen how it is possible to devise statistical tests of the fit of data to the Rasch model, either through investigating the invariance of item parameters for groups of persons or through investigating the invariance of person parameters for groups of items. Both these groups of tests, the ICC- and PCC-tests, are tests of unidimensionality but they are not equally sensitive to different deviations and a deviation that may be detected with one test may be impossible to detect with another test.

There are a number of identifiable sources of threat against the Rasch model, and it is of course of great interest to clarify which statistical tests are needed to detect different types of deviations. Such sources of deviation are discussed below.

### Item heterogeneity

Item heterogeneity, in the sense that different groups of items measure different abilities, is of course a violation of the assumption of unidimensionality.

As long as there is some basis for an a priori grouping of the items according to different hypothesized dimensions the most straightforward way to investigate this kind of deviation is to use the ML-PCC test. This is also the method to be recommended, but we shall first see if there are other methods with which item heterogeneity can be detected.

In presentations of the Rasch model (e.g. Gustafsson, 1977) it has been implied that item heterogeneity can be detected with ICCSL tests. As long as the items measuring different abilities have different discrimination parameters the ICCSL tests do in fact detect item heterogeneity, but it is of course conceivable that there are no detectable differences in the slopes of the ICC's for the different groups of items, in which case an ICCSL test would not detect multidimensionality.

That this may be the case was shown with generated data by Gustafsson and Lindblad (1978 ; cf. Brink, 1970). They demonstrated that the A-ICCSL test did not reject the Rasch model even for data generated according to an orthogonal 2-factor model, which in that case was due to the fact that every item related in the same way to a composite of the two latent variables involved. Of course, if this test and the other ICCSL tests fail to detect multidimensionality in generated data, it is also possible that they may fail to do so with empirical data.

An example will be presented to show that this is not just a highly unlikely possibility, but that it may actually happen in reality. Muthén (1978) analyzed, as an illustration of a newly developed method for factor analysis of dichotomous items, 15 items in a questionnaire assessing the personality variable internal-external locus of control (Rotter, 1966). There were data for 391 persons. The factor analysis showed that there were three lowly correlated factors among the 15 items.

The fit of these data[1] to the Rasch model has been investigated with the A-ICCSL test[2], and a very good fit was found ($\chi^2=22.4$, df=28, $p<.76$). Since there is no reason to distrust the factor analysis it seems that the A-ICCSL test in this case is not a test of the unidimensionality of the items in the questionnaire.

Additional support for this conclusion is obtained if the data are also analyzed with the ML-PCC test, with the items grouped into three scales according to their highest loading in the

factor analysis. There were 6,6 and 3 items in the scales. Applying the ML-PCC test to two scales at a time, the following results were obtained: 1 vs 2: $\chi^2$=123.8, df=35, p<.00; 1 vs 3 $\chi^2$=42.3, df=17, p<.00; 2 vs 3: $\chi^2$=60.0, df=17, p<.00. These results show clearly that the three scales measure different dimensions.
In this case we must draw the conlusion that the ICCSL-tests are not sensitive to multidimensionality among the items, and a warning must be issued, to not accept fit to the model, as shown by an ICCSL test, as evidence of unidimensionality.

To test this kind of multidimensionality in a more proper way within the framework of the Rasch model, the ML-PCC test should be used. That test, however, is a confirmatory test which requires that the items be grouped into sub-sets before any analysis is performed, and often the prior information is too weak to provide an adequate basis for this. In these cases, it does seem necessary to use factor analysis to obtain information about the dimensionality of the observations.

It is well known that factor analysis of dichotomous items has many problems, both when phi-coefficients are used (Ferguson, 1941) and when tetrachoric correlations are used (Gourlay, 1951; Lord & Novick, 1968, p. 349). Factor analytic methods specially designed for dichotomous items have, however, recently been developed (Christoffersson, 1975; Muthén, 1978). Statistically these methods are attractive but they involve great computational complexities, which at present limits their usefulness to smaller sets of items (less than 20, say; Muthén, 1978).

However, even though there are still unsolved problems in factor analysis of dichotomous items, the factor analytic methods are likely to give much information about the dimensionality and grouping of the items that is impossible to obtain in any other way. It should also be pointed out that even quite imperfect factor analytic methods can be used, since the results are checked with the ML-PCC test. Thus, for example, if a factor analysis of phi-coefficients has produced "difficulty" factors (Ferguson, 1941) these can be detected with the ML-PCC test.

## Item bias

If certain groups of items are systematically too easy or too difficult for certain sub-groups of the sample, this represents a special case of item heterogeneity which is referred to as item bias. An example of item bias may be that certain items favor the boys in a sample, while certain other items favor the girls.

Item bias can be detected in two ways. One possibility is to use the ML-PCC test, with the items grouped into internally homogeneous scales which are supposed to give different "profiles" of performance level in different groups. The other possibility is to use the A-ICC test, with the sample of persons divided into groups, such as boys and girls.

## Speededness

Speededness of the test is obviously a violation of the model assumptions since if a person does not have time to attempt an item, any statement about the probability of a correct answer as a function of ability is meaningless.

In a speeded test the items early and late in the test measure different abilities as long as "speed" and "power" are not perfectly correlated, so speededness can be detected with the ML-PCC test, if a proper grouping of the items is used.

Speedness is also possible to detect with the ICCSL tests. Persons with low raw scores do not even attempt the items late in the test, so those items will appear to have too high a discrimination (cf. Mead, 1976a, p.9).

## Guessing

If guessing takes place, which is particularly likely when multiple-choice items with few response alternatives are used, the ICC cannot be represented with one parameter only; a model like the 3-parameter model (Birnbaum, 1968) is needed to represent such data adequately.

Guessing can be detected with the ICCSL-tests if the items are of unequal difficulty and the persons have unequal ability: too many low-ability persons will answer the difficult items correctly, whereby they obtain too high a raw score, which in turn implies that on the easier items where the proportion of guesses is smaller, the low-ability persons will appear to perform too poorly. The easier items will thus appear to have too high a discrimination and the more difficult items will appear to have too low a discrimination.

Mead (1976b, p.96) showed that guessing also affects the apparent value of the person sensitivity parameters, so guessing can also be detected with the ML-PCCSL test.

Non-independence of responses

The assumption of local statistical independence implies that the response made by a person to an item must be independent of the responses to the other items in the test. This assumption can be violated in several different ways, such as by learning effects and by constrained responses. If, for example, four responses are derived from a question requiring the pairing with respect to meaning of four given English words with four given Swedish words, those of the examinees who know three of the answers will automatically have their fourth answer correct as well. Or, to take another example, if the answer given on one item affects the answer given on another item, the assumption of local statistical independence will be violated.

As has already been pointed out, the assumption of local statistical independence is equivalent to the assumption of unidimensionality, and non-independence of responses can be detected with the ML-PCC test, if the items thought to be affected by such non-independence are grouped into one group, and the other items grouped into another group.

Heterogeneous item discrimination

The ICCSL-tests are by definition sensitive to variations in the discrimination of the items, so this kind of deviation

from the Rasch model can easily be detected.

As has already been pointed out, it is, however, quite difficult to differentiate between violations of the assumption of unidimensionality and violations of the assumption of homogeneous item discrimination. This question was discussed at a rather abstract level in section 1, and here a few more comments will be made in relation to a concrete example.

Gustafsson (1977, pp. 63-69) analyzed an inductive reasoning test composed of number series items and found that two items gave evidence of too low a discrimination. The quite obvious explanation was that these items posed a much higher demand for arithmetical skills than did the other items.

The poor fit of these items was interpreted as being due to multidimensionality, which is reasonable according to any definition of unidimensionality. For example in a factor analysis the two items might define a factor of their own. However, had there been only one item of that kind the item set would have been unidimensional according to the Lord & Novick definition of unidimensionality, with a large item-specific component for the item posing high demands for arithmetical skills.

This illustrates the very blurred line of distinction between multidimensionality and heterogeneous item discrimination and that models which allow the item discriminations to vary do not easily allow generalization beyond the specific set of items analyzed.

Heterogeneous person sensitivity

Lumsden (1977, 1978) drew attention to the fact that person reliabilities may differ. If that is the case, the person sensitivity parameter in the Lumsden model (1.4) would be different for different persons, which is a violation of the assumptions of the Rasch model.

This kind of threat against the Rasch model has not been studied at all, but the possibility of varying person reliability must be taken seriously, both for practical and for theoretical reasons.

In principle,heterogeneous person sensitivity parameters can be detected with the ML-PCCSL test, but this is probably not the best way to study this kind of phenomenon. The test is likely to have a low power only, and it is sensitive to many other sources of threat as well. Furthermore, it is not likely that it will ever be possible to estimate the person sensitivity parameters, so not very much is gained by only knowing that they differ.

A better approach may be to try to find another variable, correlated with the person sensitivity parameters, and to use the A-ICC test with the sample grouped according to level of performance on this other variable. If the level of the person sensitivity parameters differs between the groups, it will be found that the item parameters are not invariant over the groups (cf. Lumsden, 1978). Such an approach would allow a more powerful test of the hypothesis, and a proxy for the person sensitivity parameters would be available. An important problem is of course what variables are likely to be related to intra-individual variability, but it does seem that personality variables are useful; Rankin (1963), for example, found that the reliability of reading tests was higher for introverts than for extraverts.

Should it be found that the person sensitivity parameters in ordinary applications do show a substantial variation, this would imply great problems from the point of view of the Rasch model, since it would not be possible to use the same model for all persons. Such a finding could be quite useful from a prediction point of view, however, within the framework of moderated regression (e.g. Ghiselli, 1965).

## Discussion

A rather long list of possible sources of deviation from the Rasch model has been compiled, and no doubt the list could

be made even longer. However, two important conclusions emerge. The first conclusion is that it is possible, in principle at least, to detect the deviations from the Rasch model, even though at times an active search is necessary. The other conclusion is that the ICCSL tests do not suffice to make a complete evaluation of fit. Nevertheless, tests sensitive to variations in the slopes of the ICC's are those that have been primarily used, and if such a test has not shown a poor fit, this has been taken as an adequate overall fit of the data. The ML-PCC test, which has never been used before, is, however, a necessary complement to such tests.

## 5. Problems in the use of statistical tests to assess fit

From the discussion above, the reader may have gained the impression that statistical tests can be used without any problems as long as they are in principle sensitive to a certain deviation. This is, of course, not so. In fact, the use of statistical tests is fraught with several problems, of which it is necessary to be aware.

Very large samples form a special source of problems. This is because no model can ever be supposed to be perfectly fitted by data, so with a sufficiently large sample any model would have to be discarded. In connection with this problem Martin-Lör (1974a) stated:

> This indicates that for large sets of data it is too destructive to let an ordinary significance test decide whether or not to accept a proposed statistical model, because, with few exceptions, we know that we shall have to reject it even without looking at the data simply because the number of observations is so large. In such cases, we need instead a quantitative measure of the size of the discrepancy between the statistical model and the observed set of data... (p,3).

Martin-Löf (1974a) derived such a measure, called redundancy, from concepts in the statistical information theory, which on an absolute scale measures the deviation between a statistical model and a set of data. This measure can thus be used when the fit of a large set of data is investigated, even though it does not seem very useful until there are tens of thousands of cases (Gustafsson, 1977, pp. 57-61), at least not for short tests

Another way to come to grips with the problems caused by large sets of data is to replace the inferential methods with descriptive methods, based on graphical descriptions of the deviations. With some experience it it thus quite easy to use the GR-ICCSL test to judge the size of magnitude of the variations in the slopes of the ICC's.

Problems are also caused by samples that are too small. Thus, the statistical tests are only asymptotically chi-square distributed, so with samples that are too small there is a risk that the test-statistic does not have the distribution assumed.

It has been argued that the A-ICC test requires a large sample to be applied with confidence (Mead, 1976b, p.34; Hambleton et al.,1977, p.63). Some preliminary simulation studies indicate, however, that the asymptotic properties of this test apply reasonably well already with as few as 50-100 persons within each group (Gustafsson, 1977, p. 54-55). The ML-ICCSL test, however, does not enjoy as good properties in this respect as does the A-ICCSL test. This is because the former test uses the results for each score group, while in the A-ICCSL test small score groups are pooled; therefore, the asymptotic properties come inte force for much smaller samples for the A-ICCSL test than for the ML-ICCSL test. It does seen wise to be cautious in interpreting the results from the ML-ICCSL test when any score group contains less than 10 observations, say.(Empty score groups cause no problems, however).

A greater problem caused by small samples is that the power of the test may be too low to detect even sizeable deviations from the Rasch model. Since the power of the tests is a function of a large number of factors, it seems impossible to give any generally valid rules for the sample sizes needed to detect deviations of different sizes. However, to give some general information about the power of the tests and to study the effects on power of different factors, some simulation studies have been performed.

## Study I: The power of the A-ICCSL test against heterogeneous item discrimination

Many of the deviations from the Rasch model appear as varying item discrimination, so it is important to have at least some rough information about the power of the ICCSL-tests against this kind of deviation. Only small samples of person will be used, so only the A-ICCSL test will be studied.

In the simulations the following factors were varied:

Number of items: 15 and 30.

Test design: One set of "peaked" tests and one set of "spaced" tests were simulated. In the peaked tests all items had a difficulty of zero at the log scale. The spaced tests contained the difficulties -2,-1,0, 1 and 2, with three items at each level of difficulty in the 15-item tests, and with 6 items at each level of difficulty in the 30-item tests.

Amount of deviation: To simulate a small amount of deviation, the discrimination parameters 0.8, 1.0 and 1.2 were used, with each discrimination parameter being represented by the same number of items at all levels of difficulty. To simulate a large deviation from the model, the discrimination parameters 0.5, 1.0 and 1.5 were used.(cf. Hambleton & Traub, 1971).

Sample size: 150 and 300.

Standard deviation (SD) of person parameters. The person parameters were sampled from two normal distributions with zero means, one with a small SD of .71 and one with a large SD of 1.22.

For each combination of levels of these factors 100 sets of data were generated according to the Birnbaum model, using the feedback shift register generator (Lewis & Payne, 1973) as the basic generator[3)]. The data were analyzed with the A-ICCSL test, with the score groups grouped in such a way that the parameters were practically always estimated within two roughly

equal-sized sub-groups. In some cases it was impossible to compute the test (cf. Gustafsson, 1977, p. 49), so for some combinations the results are based on a lower number of replications than 100.

The percentage of replications in which the p-value of the test was lower than .05 is shown in Table 1 for all the combinations of levels of the factors.

- - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - -

All the factors studied affect power. The sample size and number of items tend to influence power in the same way, at least for the peaked test, which shows that the number of responses analyzed is important.

Deviations are more easily detected in a peaked test than in a spaced test. This is because the amount of information in a response is at a maximum when the probability of a correct answer is .50 and in a spaced test there are fewer such occurrences than in a peaked test. Had the mean of the distribution of person parameters been varied as well, a lower power would have been found when the mean of ability differs from the mean difficulty of the test.

The SD of the person parameters strongly affects power. In fact, when the SD is zero, the test has no power whatsoever against this type of deviation (cf. Wright, 1977b). That this is the case is not always realized; Wood (1978),for example, reported that the Rasch model fits random data -- and seemed surprised at the finding.

When a large amount of deviation is present in the data, the test provides an adequate power in almost all instances. The most notable exception to this is when the factors combine most unfavorably, i.e. a short and spaced test, a low SD and a small sample.

When there is a small to moderate amount of deviation, the power is adequate only in the most favorable combination of levels on the factors.

One should hesitate to draw any general conclusions on the basis of a study as limited in scope as this one. It does appear, however, that as long as the SD of ability is not too low (around 1.0, say) and the difficulty of the test is adequate for the sample, a reasonable power to detect moderate heterogeneity of the item discriminations is obtained when 10 000-20 000 responses are analyzed.

Study II: The power of the A-ICCSL test against guessing

If guessing is a factor affecting performance, this tends to affect the apparent value of the discrimination parameter in the Birnbaum model.For easy items a high discrimination is observed, and for difficult items a low discrimination is observed. Some simulations have been performed to study the power of the A-ICCSL test to guard against this type of deviation from the Rasch model.

It would make only little sense in making simulations on peaked tests when guessing is the threat; the test has any power only when there is some variation of the item difficulties. Therefore, only spaced tests, designed in the same way as in Study I, were included.

Only one amount of deviation was studied: all items were supposed to have a value of .20 on the guessing parameter in the 3-parameter model (Birnbaum, 1968), and all the discrimination parameters were assumed to be unity.

Except for these changes in the design, the study was carried out in the same way as Study I, using the same levels on the other factors, except, of course, that the data were generated according to the 3-parameter model.

The results are presented in Table 2. Again, all the factors

- - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - -

affect power and they do so, of course, in the same way as was found in Study I. It is found, however, that in no case is the power adequate for the 15-item test, and only with a high SD and a sample of 300 persons is the power large enough for the 30-item test.

Comparing the figures presented for the spaced test in Table 1 with those presented in Table 2, it is found that with a guessing parameter of .20 the effect on the apparent discrimination is somewhat larger than what was labelled a small variation in the item discriminations. It would seem, however, that here too some 10 000-20 000 responses would be needed to detect presence of guessing of this amount, granted that the SD is not low and that there is a substantial variation in the item difficulties.

Study III: The power of the A-ICCSL test against both heterogeneous item discrimination and guessing

Only rarely can it be suspected that there is only one kind of deviation from the Rasch model in the data. To study the power of the A-ICCSL test against two sources of deviation, a study was performed in which both guessing and varying item discrimination was present.

The same design as in Study II was used, except that the discrimination of the items was also varied, using the discrimination parameters 0.5, 1.0 and 1.5, with each discrimination parameter being represented by the same number of items at all levels of difficulty.

The data were generated according to the 3-parameter model and again 100 replications were used.

The results are presented in Table 3. As compared with Study II the power is higher, as would be expected from the fact that another sizeable deviation has been introduced. But comparing

Insert Table 3 about here

the results with those obtained in Study I, when a large amount of deviation was simulated for a spaced test, a lower power is found when guessing is also introduced. This is because these two kinds of deviation partly cancel out: the easy items with too low a discrimination and the difficult items with too high a discrimination obtain a more "normal" discrimination as consequence of the guessing.

Examples could easily be constructed in which the effects of two deviations on the discriminations cancel out completely, resulting in no power whatsoever of the test to discover any of them. Of course, it is also possible for different deviations to work in the same direction, so that the deviations magnify each other.

## Discussion

The simulation studies presented here indicate that the A-ICCSL test should be sufficiently powerful against alternative models of the 2- and 3-parameter type if samples of 500-1 000 persons are used and if the tests contain about 20-40 items. It must be kept in mind, however, that the SD of ability is a factor critically affecting power, as is the range of item difficulties when guessing is present.

The possibility of trading relationships between different violations must be taken seriously. Using a goodness-of-fit test only, it is not possible to decide whether there are one or more deviations from the model, so this information must be taken from other sources. For some types of possible deviations this is not difficult. It should be possible to judge from the item type wheter or not a substantial amount of guessing is present, and if a test is speeded, there tends to be a large amount of omitted responses for the items late in the test. If such sources of deviation can be identified it should be seriously considered if any goodness-of-fit test should be carried out at all; it is already known that the Rasch model cannot be expected to fit the observations, and there is a risk that there will be trading relationships between those deviations, and others not so easily detected.

When the problem of too large samples was discussed, it was suggested that descriptions of deviations using graphical metods should be used. This recommendation also applies when there is a risk that the sample is too small; a deviation impossible to detect with a power-less statistical test may be possible to detect with a graphical test.

Only the power for the A-ICCSL test has been investigated here, and similar investigations could be carried out for the other tests. It is not expected, however, that very different conclusions would be arrived at. Thus, the ML-PCC test seems quite powerful when "normal" samples are used, as long as there is some variation in the abilities measured by the different groups of items.

## 6. Evaluating fit in different measurement contexts

The question of fit is of course not an absolute one and it is quite obvious that the purpose for which the model is used should decide how to treat the goodness-of-fit problem.

It does seem possible to make a distiction between two major classes of application of the Rasch model into which the goodness-of-fit problem enters differently. In the first of these the Rasch model is used as a criterion, against which characteristics of the observations themselves are evaluated. This kind of application is based on the fact that the Rasch model formalizes desirable characteristics of measurements, (i.e. unidimensionality and sufficiency of raw score as an estimator of ability, cf. Gustafsson, 1977, Wright, 1977b),and fit to the model is used to draw inferences that the observations in fact enjoy these desirable characteristics.

In the second class of applications, the Rasch model is used as an instrument to solve one or more practical measurement problems, such as linking and equating tests, carrying out tailored testing, optimizing tests, constructing item banks and so on (e.g. Wright, 1977a). In this kind of application, the solution of a practical measurement problem is the main objective, and the characteristics of the observations themselves are important only to the extent

that they help/prevent the achievement of the end.

Evaluating fit when the Rasch model is used as a criterion

It is fairly commonly accepted that in work with a theoretical orientation the scales into which observations are assembled should be homogeneous (e.g. Lord & Novick, 1968, p.351). As was pointed out by Lumsden (1976), the notion of unidimensionality has, however, been seriously neglected both by constructors of tests and by test theorists.

The unidimensionality assumption of the Rasch model, along with the availability of goodness-of-fit tests makes, in principle at least, this model useful in investigations of the unidimensionality of sets of observations.

It can be noted, though, that doubts have been expressed as to the possibility of using the Rasch model as a criterion of unidimensionality. Speaking primarily about the Rasch model and the normal ogive model, Wood (1976) stated:

> These item response models seem to be remarkably elastic concerning the motley collections of items they will fit (Wood, 1976, pp. 258-259).

And:

> It looks as if, by one means or another, heterogeneous collections of items can be made to fit response models even though inspection strongly suggests that the items are not congruent, as where groups of items call on psychologically distinguishable processes... (Wood, 1976, p. 260).

The background of these statements is almost certainly that incomplete evaluations of fit have been made. For the Rasch model only ICCSL tests have, no doubt, been employed, and it has already been shown that such tests may fail to detect even serious violations of the assumption of unidimensionality.

Thus, it is obvious that when the Rasch model is used as a criterion, high standards of fit must be set, and it is necessary that several tests which each guard against different

deviations from the model are applied. In addition to ICCSL tests, the ML-PCC test should have a central place in this kind of application, since the latter test forms a direct test of unidimensionality.

This test, however, requires that the items are grouped into subsets before it is computed, which implies that information about the dimensionality of the observations must be taken from other sources. It has already been suggested that factor analysis is useful in this context, but also information derived from a careful scrutiny of the items and observations of solution processes, are likely to be useful (cf. Cronbach, 1970, pp. 474-475).

When the Rasch model is used as a criterion, the power of the statistical tests is essential. Whenever it is suspected that the power is insufficient, a closer look at the problem should be taken, perhaps through conducting a specially designed simulation study in which the characteristics of that particular situation are represented.

The use of the Rasch model as a criterion is above all of interest in work with a theoretical orientation. This implies that the items must be constructed from theoretical starting points, and these theoretical notions should also direct the evaluation of fit. In such work the Rasch model is also likely to prove useful to test specific hypotheses about test items, without it being regarded a failure if the model is rejected.

It is true that most test construction is essentially atheoretical and, as has been pointed out by Levy (1973), there is only a weak relation between test theory and psychological theory:

> Statistical manipulation of test results is sometimes used as a poor substitute for operational control of item content and format at the test development stage. Much needed are tests constructed to test hypotheses, and fewer hypotheses about tests (Levy, 1973, p. 37).

This state of affairs is not likely to change as a function of adoption of the Rasch model. Should, however, a greater theoretical sophistication come about among test constructors, it is

likely that the Rasch model will be found to contribute to their work.

Whitely and Dawis (1974) spoke in a similar vein:

> "...the lack of impact of the Rasch model in test development is due more to the current status of trait measurement than to the properties of the model.(p.77).

Evaluating fit when the Rasch model is used as an instrument

The Rasch model can be used as an instrument to solve a range of practical measurement problems (e.g. Wright, 1977a). Here too, the fit of the data to the model is important, but the question of fit is nevertheless subordinated to the solution of concrete measurement problems. This implies that lower standards of fit can sometimes be set, that all possible deviations from the model assumptions need not necessarily be considered, and that in fact large deviations in the data from the model assumptions can sometimes be tolerated.

If it is known that a set of data fit the Rasch model, it follows from the mathematical structure of the model itself that it can be used to solve practical measurement problems. The reason, however, why the Rasch model sometimes might be used as an instrument, in spite of poor fit, is that deviations from the model do not necessarily jeopardize applications. Unfortunately, very little is known about the robustness of the Rasch model against different types of deviations for different types of applications, and this is an area where much research is needed.

Some research has been carried out, though, and it may be instructive to consider some of that in greater detail to see how the goodness-of-fit problem can be handled when the Rasch model is used as an instrument.

The area of application where most research on the robustness of the Rasch model has been carried out is on the equating of tests, i.e. expressing on the same scale raw scores obtained on different tests. In principle, this problem is easily solved with the Rasch model through first estimating the item parameters

for the two tests on a common scale, and then deriving the ability scales which specify the conversion of raw scores into estimates of ability (e.g. Wright, 1977a; Rentz & Bashaw, 1977).

It has been shown (e.g. Wright, 1968; Whitely & Dawis, 1974) that estimates of the mean of ability for a group derived from easy and difficulty items in a test come quite close. In those studies the data did not fit the model, which indicates that the estimates of ability are quite robust against deviations from the model.

However, Slinde and Linn (1978) argued that it should also be possible in vertical equating of tests (i.e. equating tests of different difficulty) to use the item parameters estimated in any group of persons to estimate the abilities in any other group of persons. They compared the means of ability estimates obtained from easy and difficult tests for groups of different levels of ability, using item parameters estimated either within the same group of persons, or estimated within a group of persons of another level of ability. It was found that a reasonably good vertical equating could be achieved when the item-parameters estimated within the groups were used, but not when item-parameters estimated within another group were used. On the basis of these results, Slinde and Linn (1978) questioned the usefulness of the Rasch model in solving the problem of vertical equating of tests.

It should be pointed out that a partial explanation of the poor results obtained by Slinde and Linn (1978) is that they used an illegal grouping of the sample into levels of ability; they used performance on a subset of the items only as the basis for the grouping, a procedure which introduces a spurious lack of fit even when the data fit the model (Gustafsson, 1979b). However, Slinde and Linn (1979) have presented another study which allowed very much the same conclusion.

The Slinde and Linn requirement that it should be possible to use the estimates of parameters from any group of persons is a reasonable one, since in some cases this is necessary in equating tests. It does seem rash, however, to draw a general

conclusion about the inability of the Rasch model to solve the problem of vertical equating on the basis of a few empirical studies alone, and without supplying any reasons for the failure.

Slinde and Linn (1978, 1979) suggested that an LT-model which allows the slopes of the ICC's to be different might be needed to solve the problem of vertical equating. Of course, in the presence of heterogeneous item discriminations the item parameters will always differ when estimated within groups of different level of performance, but depending upon the exact the kind of violation of the assumption of homogeneous item discrimination, the biasing effects in vertical equating will be different.

Some simple simulation studies have been performed to illustrate this. Data were generated to follow the Birnbaum model for three tests with 60 items in each, 30 of which had the difficulty -1 and 30 of which had the difficulty 1. In one of the tests the item discriminations were not correlated with difficulty, there being 10 items each with discrimination parameters 0.5, 1.0 and 1.5 at each of the levels of difficulty. This test will be referred to as the ZCORR test. In another test there were 10 items with discrimination 1.0 and 20 items with discrimination 1.5 among the easy items; among the difficult items there were 10 items with discrimination 1.0 and 20 items with discrimination 0.5. This test will be referred to as the NCORR test, since it simulates the case when there is a negative correlation between discrimination and difficulty, such as is the case when the test items allow guessing. Finally, in the third test (PCORR) the frequencies of items with high and low discriminations were reversed at the two levels of difficulty as compared with the NCORR test, to simulate a test with a positive correlation between discrimination and difficulty, such as tends to be the case for a speeded test.

For each of these three tests data were generated for 1 000 persons, with the ability parameters being sampled from a normal distribution with zero mean and unit standard deviation. Persons with a score equal to 30 or lower formed a "low" group, and the rest of the sample formed a "high" group. The item parameters

for the total set of 60 items were estimated within the two groups, and the mean of ability for each group was estimated separately for the easy and difficult items, using the item parameters estimated both within the same group and the other group of persons.

Table 4 presents, for the easy and difficult items separately,

- - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - -

the difference between the means obtained when using the item parameters estimated within the same group and those estimated within the other group. It could of course be argued that the differences between means obtained on easy and difficult items should be presented instead, since the problem of vertical equating is studied. In this case however, these are not directly comparable, since some of the persons in the high group had a perfect score on the easy items and some of the persons in the low group had a zero score on the difficult items.

The figures presented in Table 4 show that for the ZCORR test only small differences are found when the estimates of ability are based on item parameters estimated within groups of different levels of ability. For the PCORR and NCORR tests, however, there is a large bias, with the direction of bias being different depending upon the sign of the correlation between item difficulty and item discrimination.

Using figures presented by Slinde and Linn (1979), the corresponding differences have been computed for that study. The pattern of differences found coincides with that found for the NCORR test, as might be expected from the fact that the test analyzed by Slinde and Linn (1979) was a multiple-choice test heavily influenced by guessing.

This brief analysis thus makes it likely that the negative conclusions drawn by Slinde and Linn as to the possibility of using the Rasch model as an instrument in the vertical equating

of tests was due to a negative correlation between item difficulty and item discrimination in that study. Had a test with the same amount of deviation but with a zero correlation with difficulty and discrimination been analyzed, a much more positive conclusion would have been arrived at.

It must be stressed that this analysis of the robustness of the Rasch model is very limited in scope and allows very limited generalizations only. Thus, attention has been confined to the estimates of the mean of ability for groups of persons, but it is well known that in the presence of heterogeneous item discrimination the Rasch model is less efficient than other LT models (Hambleton & Traub, 1971; Reckase, 1978). Parenthetically, it should also be pointed out that Andersen and Madsen (1977) have recently presented a superior solution to the problem of estimating the parameters of the latent population distribution. The robustness of that method against deviations from the Rasch model assumptions remains as yet to be studied.

The purpose of this digression has been to show that the Rasch model sometimes is quite robust against deviations from the model assumptions, while at other times it is not robust at all. This suggests that when the Rasch model is to be used as an instrument on data not fitting the model, the deviations from the model should first be analyzed and described, and it should then be investigated whether the model is robust against these deviations for the particular application intended.

Of course, the Rasch model is best used as an instrument when the data fit the model. It should therefore also always be investigated if it is possible to obtain fit of data to the model. Strategies for doing this are discussed in the next section.

## 7. Obtaining fit of data to the Rasch model

It does appear that, on the whole, a rather simple strategy is followed to obtain fit of data to the Rasch model. This standard procedure may be described in the following, somewhat simplified, way: a set of items is given to a sample of persons and an overall ICCSL test is computed. If this test is significant,

which is usually the case, the p-value of fit to the model of each item is computed, or a graphic test of item fit is made, and those items which do not fit are excluded. A new overall ICCSL test is then computed, usually with the same sample of persons, and unless a non-significant value on the test statistic is obtained, the process is carried out again, excluding more items, until a reasonably good overall fit is obtained.

It is submitted here that this strategy is likely to result in a spurious fit only, and that it should only rarely be used. In view of current practice this is a strong assertion, but several reasons can be cited in support of it.

One reason is of course, as has been shown above, that the ICCSL tests represent only a partial evaluation of fit to the model, and they can fail to detect even very serious deviations from the Rasch model. Other tests, and above all the ML-PCC test, should therefore also be used to study item heterogeneity.

Another reason why the strategy based on exclusion of items should not be used is that there may be trading relationships between different violations of the model assumptions, as was shown in Study III in section 5. Consider for example a slightly speeded multiple-choice test with heterogeneous items. Speededness and guessing tend to affect the discriminations in opposite directions and item heterogeneity may also affect the discriminations. It is very likely that a large proportion of the items in such a test which do show a good fit do this because the effects of the different violations cancel out. If "poor-fitting" items are excluded, a good overall fit, as evidenced by an ICCSL test will eventually be obtained, but that good fit has been obtained through capitalizing on such trading relationships, and on chance effects. When this kind of "fit" has been obtained, the implications which are otherwise associated with fit of data to the Rasch model do not hold true.

A third reason why items should not be routinely exclued is that there may be deviations from the model where other steps should be taken to obtain fit. If, for example, the main reason for the

poor fit of a set of items is that the examinees have been given too short a testing time, the best way to obtain fit is to give the test with a more liberal time limit. Or, to take another example, if the test consists of multiple-choice items with a few response alternatives on which the subjects have been given the instruction of guess if they do not know the correct answer, it does not seem wise to select those items which appear to fit the model in spite of the guessing; instead the opportunities to guess at all should be minimized if the Rasch model is to be used.

But there is also a fourth, and even more important reason why development of Rasch scales on the basis of exclusion of poor-fitting items cannot be recommended as a general strategy. This is because tests of the fit of single items, in the presence of gross deviations from the model, are in principle illogical: the basic requirement of the Rasch model is that the items shall be homogeneous, so what is tested is, in fact, if the items fit with each other, not if they fit the model. If tests of item fit indicate that just a few of the items do not fit, this can of course be interpreted as showing that these items do not fit with the other items, and hence not the model. But if a larger proportion of the items show misfit, the item set is so heterogeneous that there may be subset of items in the set, each of which shows a good fit to the model, but which do not fit with each other.

Suppose for example, that a set of items all measure the same ability but that they have different discrimination parameters (which is, of course, a highly hypothetical situation). If items are excluded on the basis of tests of item fit, those items will be retained which have an intermediate level of discrimination. But there is no assumption in the Rasch model which says that items shall have an intermediate level of discrimination; all that is required is that the items shall be homogeneous with respect to discrimination. Thus it may well be possible to select a subset of highly discriminating items which fit the model. If the scale is to be used to measure individual differences, such a scale composed of highly discriminating itmes will have better properties than a scale composed of

items with an intermediate discrimination, at least if the discrimination is not so high, and the difficulties not so uniform that the attentenuation paradox appears (Loevinger, 1954).

In passing, it can be noted that studies have been carried out (e.g. Tinsley & Davis, 1972) in which the Rasch model has been compared with other methods for item screening. In these studies the tests of item fit have been used to select items for the Rasch scales, and it has not been realized that items which appear to have too high a discrimination could have been selected just as well.

Other examples where tests of the fit of single items may give absurd results are easily envisaged. If, for example, a set of items is heterogeneous in the sense that two dimensions are covered, an ICCSL test may, but need not, indicate a poor fit. But if a process of item selection is carried out we will end up, at best, with a scale covering only one of the dimensions in the original set. What should be done in such a case is of course to sort the items into internally homogeneous subsets each of which will show a good fit to the model.

Gustafsson and Lindblad (1978) presented an empirical example of that situation. In analyses of a test of English grammar for Swedish students it was found that a set of items measuring knowledge of irregular verbs did not fit the model. But in a separate analysis of these items it was found that they did fit the model, as did the rest of the items, after some poorly constructed items had been excluded. Had the items measuring knowledge of irregular verbs been excluded, that would have implied an undue narrowing of the scope of the test, but through forming two scales instead of one, both kinds of items were retained.

The Rasch model has been critisized by several authors (e.g Goldstein & Blinkhorn, 1977; Whitely, 1977; Wood, 1978) because it has been thought that the strong assumptions of the model make it necessary to exclude items not fitting the model. Wood (1978), for example, said:

> By narrowing the scope of the tests in order to fit the Rasch model, we may run the risk of throwing out the baby with the bath water, even though the measurements have desirable, perhaps even necessary, properties....(Wood, 1978, p. 31).

This criticism is warranted if it is assumed that only one scale is to be used, but not otherwise; any degree of heterogeneity can be represented with the Rasch model as long as several different scales are constructed (cf. Lumsden, 1976, p. 267).

From the list of problems associated with the exclusion of poor-fitting items to obtain fit, the skeleton of an alternative strategy can be outlined. First of all the likely causes of the poor fit should be identified. If among the likely sources of deviation there are factors other than item heterogeneity, the proper actions should be taken to remove those threats against the model (i.e. remove speededness, guessing and so on). It should then be investigated if the item heterogeneity is so severe that the items should be grouped into homogeneous subsets, or if a few poorly constructed items can be excluded to obtain fit. In the next step, any suggested scale should be cross-validated on another sample of persons with further items.

In order for such a strategy to be successful,a very good kowledge of the sample, the testing situation and the content of the items is necessary; otherwise it will be impossible to trace the different sources of deviation and to group the items. The goodness-of-fit tests are likely to contribute in the evaluation of fit, but they can certainly not replace subject matter knowledge.

## Concluding remarks

If anything, it should stand clear from the discussions in this paper that it is difficult both to evaluate and to obtain fit of data to the Rasch model. It can only be hoped that this does not detract users from the Rasch model, because if used properly there are sometimes large theoretical and practical gains to be made, and especially so if the goodness-of-fit problem is given due attention.

Closely associated with the Rasch model is the theory of specific objectivity (Rasch, 1960, 1961, 1977) which says that it should be possible to compare objects (persons) independently of agents (items) and agents independently of objects. When data fit the Rasch model specifically objective comparisons of items can be made, as well as specifically objective comparisons of persons. But users of the Rasch model must bear in mind the following caution, made by Rasch himself:

> In an empirical science specific objectivity can never be fully ascertained if the objects and/or agents is an infinite set; it can only be set up as a working hypothesis which has got to be carefully tested, e.g. by exposing an extensive body of objects to a wide range of agents and analyzing the reactions. And whenever additional data are collected we must be ready to do it over again -- possibly having to revise previous optimistic conclusions. (Rasch, 1965, p. 8, with some changes of notation).

FOOTNOTES

1) I want to thank Bengt Muthén for kindly giving me access to these data, and also those persons acknowledged by Muthén (1978) for having originally contributed the data.

2) These computations, and all others reported in this paper, were made with a FORTRAN IV computer program (PML3), written by the present author for use on IBM 360/370. PML3 computes the CML estimates of the item parameters, and estimates of the person parameters. The program also computes all the goodness-of-fit tests presented here, except for tests of person fit. A copy of the program written on tape may be obtained at cost from Jan-Eric Gustafsson, Institute of Education, University of Göteborg, Fack, S-431 20 MÖLNDAL, Sweden.

3) I want to thank Philip Ramsey, now at the City University of New York, for putting into my hands this exellent random number generator.

Table 1

Percentage of successful replications in which the A-ICCSL test rejected the Rasch model at the 5 percent level in the presence of heterogeneous item discrimination.

SMALL AMOUNT OF DEVIATION

Test design

| | Peaked | | | | Spaced | | | |
|---|---|---|---|---|---|---|---|---|
| | Number of items | | | | Number of items | | | |
| | 15 | | 30 | | 15 | | 30 | |
| | Sample size | | Sample size | | Sample size | | Sample size | |
| SD | 150 | 300 | 150 | 300 | 150 | 300 | 150 | 300 |
| Low | 11 | 15 | 15 | 44 | 6 | 15 | 11 | 21 |
| High | 26 | 51 | 44 | 80 | 13 | 33 | 22 | 60 |

LARGE AMOUNT OF DEVIATION

Test design

| | Peaked | | | | Spaced | | | |
|---|---|---|---|---|---|---|---|---|
| | Number of items | | | | Number of items | | | |
| | 15 | | 30 | | 15 | | 30 | |
| | Sample size | | Sample size | | Sample size | | Sample size | |
| SD | 150 | 300 | 150 | 300 | 150 | 300 | 150 | 300 |
| Low | 53 | 89 | 89 | 99 | 28 | 67 | 57 | 64 |
| High | 99 | 100 | 100 | 100 | 81 | 98 | 97 | 100 |

Table 2

Percentage of successful replications in which the A-ICCSL test reject the Rasch model at the 5 percent level in the presence of guessing.

| | Number of items | | | |
|---|---|---|---|---|
| | 15 | | 30 | |
| | Sample size | | Sample size | |
| SD | 150 | 300 | 150 | 300 |
| Low | 12 | 16 | 15 | 32 |
| High | 26 | 48 | 59 | 96 |

Table 3

Percentage of successful replications in which the A-ICCSL test rejected the Rasch model at the 5 percent level in the presence of guessing and varying item discrimination.

| | Number of items | | | |
|---|---|---|---|---|
| | 15 | | 30 | |
| | Sample size | | Sample size | |
| SD | 150 | 300 | 150 | 300 |
| Low | 10 | 35 | 32 | 77 |
| High | 53 | 94 | 90 | 100 |

Table 4

Differences between estimates of means of ability using parameters estimated within the same group of persons and parameters estimates within the other group of persons

| | TEST | | | |
|---|---|---|---|---|
| | ZCORR | PCORR | NCORR | Slinde & Linn (1979) |
| Easy items | | | | |
| Low | -.02 | -.45 | .39 | .62 |
| High | .14 | .50 | -.36 | -.30 |
| Difficult items | | | | |
| Low | -.13 | .40 | -.44 | -.48 |
| High | .02 | -.44 | .40 | .62 |

REFERENCES


Andersen, E.B. (1973a) Conditional inference and models for measuring. Copenhagen: Mentalhygiejnisk Forlag.

Andersen, E.B. (1973b) A goodness of fit test for the Rasch model. Psychometrika, 38, 123-140.

Andersen, E.B., & Madsen, M. (1977) Estimating the parameters of the latent population distribution. Psychometrika, 42, 357-374.

Baker, F.B. (1977) Advances in item analysis. Review of Educational Research, 47, 151-178.

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, & M.R. Novick, Statistical theories of mental test scores. Reading: Addison-Wesley.

Brink, N.E. (1970) Characteristics of Rasch's Logistic model. Paper presented at the Annual Meeting of the American Educational Research Association. ED042804.

Christoffersson, A. (1975) Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Cronbach, L.J. (1970) Test validation. In R.L. Thorndike (Ed) Educational Measurement (2nd ed). Washington: American Council on Education.

Ferguson, G.A. (1941) The factorial interpretation of test difficulty. Psychometrika, 6, 323-329.

Fischer, G.H. (1974) Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen. Bern: Huber.

Ghiselli, E.E. (1965) Moderating effects and differential reliability and validity. Journal of Applied Psychology, 47, 81-86.

Goldstein, H., & Blinkhorn, S. (1977) Monitoring educational standards - an inappropriate model. Bulletin of the British Pshychological Society, 30, 309-311.

Gourlay, N. (1951) Difficulty factors arising from the use of tetrachoric correlations in factor analysis. British Journal of Psychology, Statistical Section, 4, 65-76.

Gustafsson, J.-E. (1977) The Rasch model for dichotomous items: Theory, applications and a computer program. Reports from the Institute of Education, University of Göteborg, no.63. ED 154018.

Gustafsson, J.-E. (1979) A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. Article in review.

Gustafsson, J.-E (1979b) The Rasch model in vertical equating of tests: A critique of Slinde and Linn. Journal of Educational Measurement, in press.

Gustafsson, J.-E., & Lindblad, T. (1978) The Rasch model for dichotomous items: A solution of the conditional estimation problem for long tests and some thoughts about item screening procedures. Reports from the Institute of Education, University of Göteborg, no.67.

Hambleton, R.K., & Cook, L.L. (1977) Latent trait models and their use in analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Hambleton, R.K., & Traub, R.E. (1971) Information curves and efficieny of three logistic test models. British Journal of Mathematical and Statistical Psychology, 24, 273-281.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. (1977) Developments in latent trait theory: A review of models, technical issues, and applications. Paper presented at a joint meeting of NCME and AERA in New York.

Leunbach, G. (1976) A probabilistic measurement model for assessing whether two tests measure the same personal factor. Reports from the Danish Institute for Educational Research no. 19, 1976.

Levy, P. (1973) On the relation between test theory and psychology. In P. Kline (Ed) New approaches in psychological measurement. London: Wiley, pp. 1-42.

Lewis, T.G. & Payne, W.H. (1973) Generalized feedback shift register pseudo random number algorithm. Journal of the Association for Computing Machinery, 20, 456-468.

Lord, F.M., & Novick, M.R. (1968) Statistical theories of mental test scores. Reading: Addison-Wesley.

Lumsden, J. (1976) Test theory. In M.R. Rosenzweig & L.W. Porter (Eds) Annual Review of Psychology, 27. Palo Alto: Annual Reviews Inc.

Lumsden, J. (1977) Person reliability. Applied Psychological Measurement, 1, 477-482.

Lumsden, J. (1978) Tests are perfectly reliable. British Journal of Statistical and Mathematical Psychology, 31, 19-26.

Martin-Löf, P. (1973) Statistiska modeller. Anteckningar från seminarier läsåret 1969-70 utarbetade av Rolf Sundberg. 2:a uppl. (Statistical models. Notes from seminars 1969-70 by Rolf Sundberg. 2nd ed.) Institutet för försäkringsmatematik och matematisk statistik vid Stockholms universitet.

Martin-Löf, P. (1974a) The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. Scandinavian Journal of Statistics, 1, 3-18.

Martin-Löf, P. (1974b) Exact tests, confidence regions and estimates. Pp. 121-138 in Proceedings of the Conference on Foundational Questions in Statistical Inference, Aarhus May 7-12, 1973. Memoirs No 1,Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus.

Mead, R. (1976a) Assessing the fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Mead, R. (1976b) Assessment of fit of data to the Rasch model through analysis of residuals. Unpublished doctoral dissertation, University of Chicago.

Muthén, B. (1978) Contributions to factor analysis of dichotomous variables. Psychometrika, in press.

Rankin , E.F. Jr.,(1963) Reading test reliability and validity as a function of introversion-extraversion. Journal of Developmental Reading, 6, 106-117.

Rasch, G. (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: The Danish Institue for Educational Research.

Rasch, G. (1961) On general laws and the meaning of measurement in psychology. In Proceedings of the fourth Berkely symposium on mathematical statistics. Berkeley: University of California Press, pp. 321-334.

Rasch, G. (1965) On objectivity and models for measuring. Lecture notes edited by Jon Stene.

Rasch, G. (1966) An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Rasch, G. (1977) Begrundelse for og konsekvenser af et krav om specific objektivitet. (Background to and consequences of a requirement of specific objectivity). The Institute for Educational Research, 1977.43.

Reckase, M.D. (1978) A comparison of the one-and three-parameter logistic models for item calibarion. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto.

Rentz, R.R., & Bashaw, W.L. (1977) National reference scale for reading-- Application of Rasch model. Journal of Educational Measurement, 14, 161-179.

Rotter, J.B. (1966) Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 80, (Whole no.609).

Slinde, J.A., & Linn, R.L. (1978) An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 15, 23-35.

Slinde, R.L.,& Linn , R.L. (1979) A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, in press.

Scheffé, H. (1959) The analysis of variance,New York: Wiley

Tinsley, H.E.A., & Dawis, R.V. (1972) A comparison of the Rasch item probability with three common item characteristics as criteria for item selection. Minnesota Univ., Minneapolis. Center for the Study of Organizational Performance and Human Effectiveness. ED 068516.

Wood, R. (1976) Trait measurement and item banks. In D.N.M. de Grujter, L.J.,Th. van der Kemp & H.F. Crombag (Eds): Advances in psychological and educational measurement. Londond: Wiley, pp. 247-263.

Wood, R. (1978) Fitting the Rasch model - A heady tale. British Journal of Mathematical and Statistical Psychology, 31, 27-32.

Whitely, S.E. (1977) Models, meanings and misunderstandings: Some issues in applying Rasch's theory. Journal of Educational Measurement, 14, 227-235.

Whitely, S.E., & Dawis, R.V. (1974) The nature of objectivity with the Rasch model. Journal of Educational Measurement, 11, 163-178.

Wright, B.D. (1968) Sample-free test calibaration and person measurment. In Proceedings of the 1967 Invitational Conference on Testing problems.Pinceton, New Jersey, pp. 85-101.

Wright, B.D. (1977a) Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-115.

Wright, B.D. (1977b) Misunderstanding the Rasch model. Journal of Educational Measurement, 14, 219-225.

Wright,B.D.,& Douglas, G.A. (1975) Better procedures for sample-free item analysis. Research Memorandum no.20, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B.D., & Douglas, G.A. (1977) Conditional versus unconditional procedures for sample-free item analysis. Educational and Psychological Measurement, 37, 47-60.

Wright, B.D., & Panchapakesan, N. (1969) A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.